Supplementary Material
LumiCtrl: Learning Illuminant Prompts for Lighting Control in Personalized Text-to-Image Models

Muhammad Atif Butt[1,3], Kai Wang[1,2,4*], Javier Vazquez-Corral[1,3], Joost Van De Weijer[1,3]
[1] Computer Vision Center, Spain [2] City University of Hong Kong
[3] Computer Sciences Department, Universitat Autònoma de Barcelona, Spain
[4] Program of Computer Science, City University of Hong Kong (Dongguan)
{mabutt, kwang, jvazquez, joost}@cvc.uab.es

1. Learning Illuminants with T2I Personalization Methods

Text-to-image (T2I) diffusion models allow users to synthesize objects by incorporating linguistic descriptions into text prompts, such as "a cat sitting on the mountain" or "a fashion model walking on the ramp". Although these T2I diffusion models have demonstrated remarkable abilities in generating images from text prompts, they struggle with synthesizing objects in precisely user-guided illuminations.

Recently, T2I personalization methods have been proposed, including Textual Inversion [5], Dreambooth [9], Custom Diffusion [7], and Break-a-Scene [1]. These personalization methods allow users to learn personal concepts such as friends, pets, or specific items given 4-5 images. We employ these methods to learn concepts such as cats and dogs and evaluate the performance in terms of illuminating the newly learned concept under different illuminations using text prompts: a photo of [v] dog under tungsten/fluorescent illumination. The results demonstrated in our main paper show that although these methods efficiently preserve the identity of the given concept, they struggle to illuminate concepts under text-guided illumination.

In particular, Custom Diffusion, Dreambooth, and Break-a-scene tend to synthesize concepts in a similar posture and adopt daylight illumination given in the example images. This behavior stems from the entanglement of object identity with scene illumination during fine-tuning—the model learns to associate the concept with its original lighting conditions, making it resistant to illuminant changes at inference time. Textual Inversion, on the other hand, does not synthesize the concept faithfully and fails to adopt the texture from training examples. Resultantly, these methods do not synthesize concepts under text-guided illumination, as evident in Fig. 1.

*Corresponding author.

We also analyze the performance of image editing methods, i.e., Pix2Pix [2] and IC-Light [10], showing results in Fig. 1 and Fig. 2. We employ pre-trained baselines and generate images given input image and text prompt. It can be observed that Pix2Pix and IC-Light struggle to synthesize concepts under text-guided illumination. Though IC-Light demonstrates considerably better performance in manipulating lighting direction, it has two key limitations. First, it does not preserve the spatial background information of the given image, often introducing inconsistent geometry or artifacts. Second, it struggles to understand and adapt to text-guided daylight illuminants and Kelvin temperatures, as these terms lack semantic grounding in the underlying text encoder—a fundamental limitation we identify and address with LumiCtrl.

2. Experiments

2.1. Training Examples

In this work, we curated a small set of 20 concepts to learn illuminant prompts from *Real Images*, and *Generated Images*, shown in Fig. 3. For real images, we pick images of pets including cats and dogs from multiple internet sources [3, 4, 6, 8], and we used different text prompts to generate training concepts, including llama, rabbit, and cows. We provide a list of text prompts used to generate these images in Section 2.2.

2.2. Training Prompts

We generate training examples with text prompts as below.

- a realistic photo of a walking Siberian Husky in the grass field.
- a photo of a cute realistic llama, highly detailed, cinematic illuminant.
- a photo of a white cow walking through the grass field.

Figure 1. Analyzing the capability of T2I personalization methods, and Image Editing Methods in illuminating concepts using text-prompts. **T2I Personalization Methods**: Textual Inversion, Dreambooth, and Custom Diffusion predominantly preserve lighting from training examples, failing to generate concepts under various illuminants. **Image Editing Methods**: These methods fails to photo-realistically change the illumination aligning with the text prompt. Note that, the IC-Light generations are from foreground conditioning.

Figure 2. Inferences with IC-Light and Instruct Pix2Pix. IC-Light struggles with preserving the spatial background information. Moreover, both models fail to understand the text-guided daylight illuminants and kelvin temperature. IC-Light examples are from foreground conditioning.

Figure 3. Data Samples — Concepts used in the qualitative and quantitative study.

- a photo of a heartwarming adorable rabbit sitting in the meadows.
- a photo of a white horse standing in front of a house.
- a photo of a chicken playing in the garden.
- a photo of a white sheep on the mountain.
- a photo of a rabbit in the garden.
- a photo of a cat sitting by the window in a room.
- a photo of a dog sitting on the floor by the window in an ultra realistic modern room.

2.3. Evaluation Settings

We optimized the new text tokens with multiple training prompt examples, listed below.

- a photo of $[v]$ concept captured in $[c_*]$ illumination.
- a photo of $[v]$ concept in $[c_*]$ illumination.
- $[v]$ concept in $[c_*]$ illumination.
- a photo of $[v]$ concept taken in $[c_*]$ illumination.
- a $[c_*]$ illuminated photo of $[v]$ concept.
- a photo of $[v]$ concept with [bg] background captured in $[c_*]$ illumination.
- a photo of $[v]$ concept with [bg] background taken in $[c_*]$ illumination.
- a $[c_*]$ illuminated photo of a $[v]$ concept with [bg] background.

Here we use the aforementioned set of instance prompts,

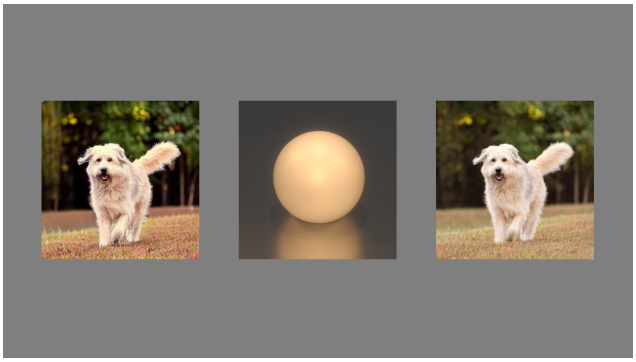

Figure 4. Setup of our User Preference Study: The monitor's background was set to a neutral gray, and participants were asked to choose which of the two images — left or right — best matched the illuminant color, based on the color displayed in the sphere in the central image.

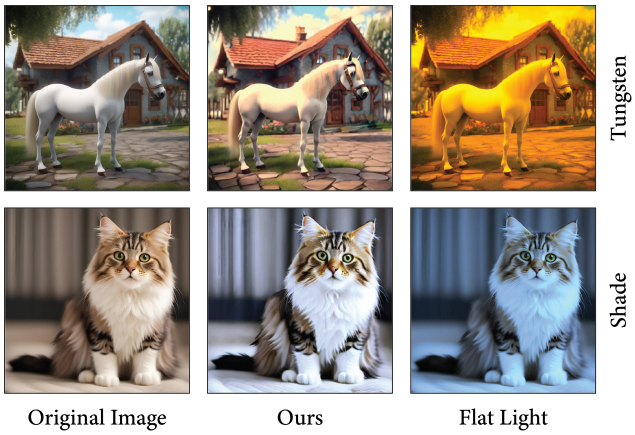

Figure 5. Comparison between our approach and the Flat Light Adaptation. Flat Light Adaptation leads to unrealistic results, while our method provides pleasant images.

where we learn embedding of the concept such as dog or cat in $[v]$ text-token, [bg] as background, and [c1],[c2],[c3],[c4],[c5],[c6], and [c7] text-tokens to learn illumination embeddings of tungsten, fluorescent, cloudy, and shade, and the three intermediate illuminants respectively.

2.4. User Study

The room in our user study was completely dark, with the monitor set to sRGB mode. The only light source during the experiment was the monitor. Participants were advised to sit about 60 cm away from the monitor, providing a 7-degree visual angle. The monitor background was set to the neutral gray, displaying a central image containing an sphere representing the light color. On either side of this image, we randomly presented results from our method and a competing one. The participant's task was to choose which of the two images best matched the prompt based on the illuminant color in the central image. Fig.4 shows an example of what the observer saw. A total of 15 participants participated in the study, none of the authors involved. The central image presents a gray sphere illuminated by the color of illuminant name —i.e. how a sphere will look under that illuminant name. To left and right, we show results of our method and one of the competing methods, in randomized order.

2.5. Comparisons versus Flat Light Algorithms

Figure 5 compares our method and a traditional illuminant adjustment based on Von Kries multiplication—*Flat Light adaptation*. We can see how the *contextual adaptation* presented in this paper generates a pleasant image. In contrast, the *Flat Light adaptation* result is unrealistic, reducing the hue diversity in the scene (the top image becomes orangish, while the bottom image becomes bluish). In particular, we

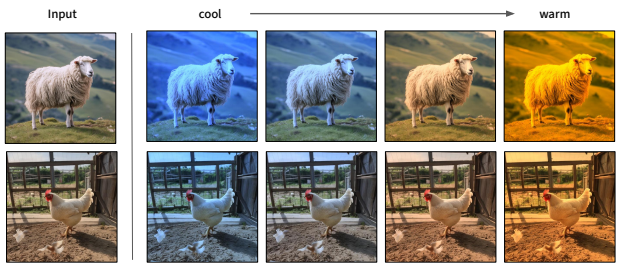

Figure 6. Demonstration of illuminating concepts into different illuminant conditions using Flat Light Adaptation. It can be noted that image becomes more unrealistic, when the illuminant is scaled higher in both— the cool, and warm conditions. Secondly, Flat Light Adaptation cannot create illuminance adaptive shadow, which is one of the main drawbacks of this approach.

present a demonstration of the illuminating concepts under different illuminant conditions using the flat light adaptation in Fig. 6. It can be noted that the image becomes more unrealistic, i.e., extreme bluish or extreme orangish when the illuminant is scaled higher towards cool and warm conditions, respectively. Secondly, the flat light adaptation cannot create illuminance adaptive shadow, which is one of the main drawbacks of this approach. For instance, the shadow of the concepts is exactly same across all the illuminants which makes it unnatural, as the intensity of shadow of the concept scales with the changing daylight illumination in the real world. However, *LumiCtrl* generates illuminance adaptive shadows as shown in Fig. 7.

2.6. Qualitative Results

We provide additional qualitative examples demonstrating the illumination of both—the real and T2I generated images into seven real-world daylight illuminations. The re-

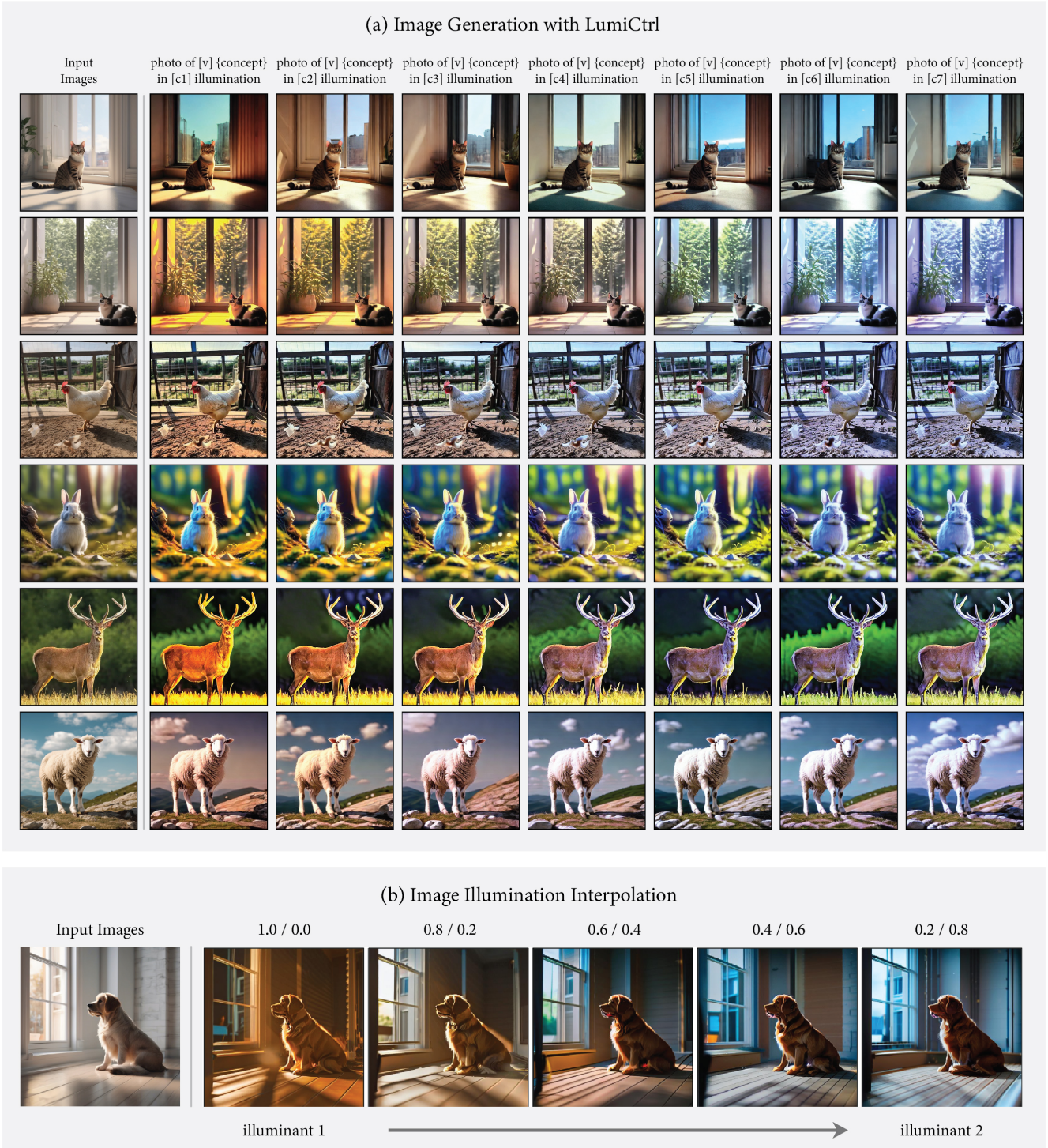

Figure 7. Additional qualitative results of *LumiCtrl* illuminating real and T2I generated concepts given text prompts. The {concept} in the prompt represents the name of the concept in the given images, such as cat, dog, rabbit, deer, and sheep.

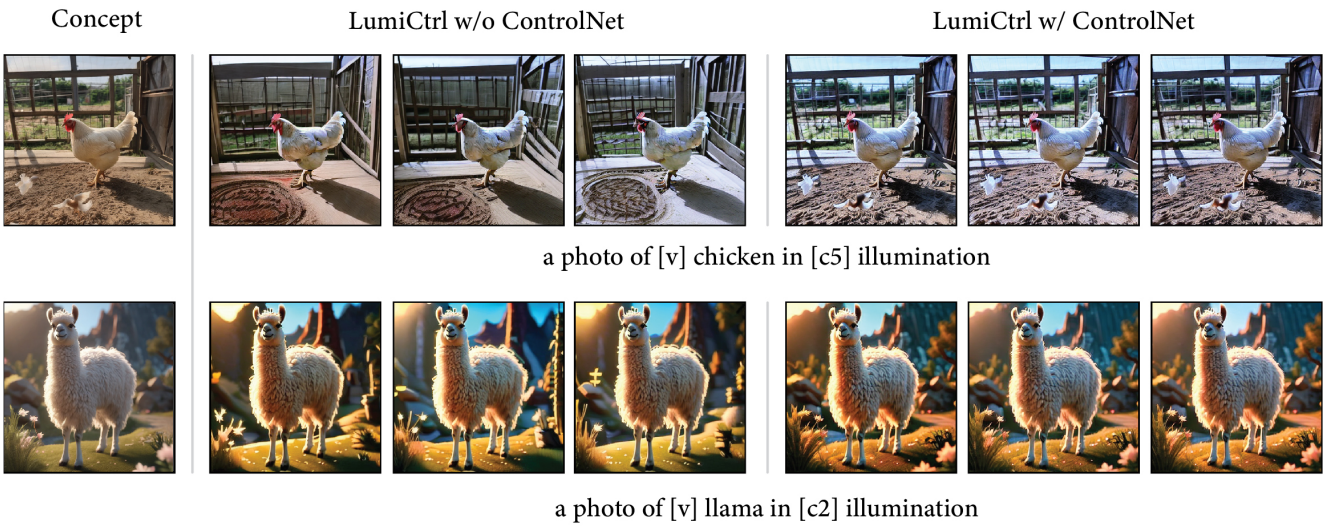

Figure 8. Demonstrating the comparison between the *LumiCtrl*'s pipeline with and without Edge-Guided Prompt Disentanglement.

sults are presented in Figure 7 which shows the capabilities of *LumiCtrl* in illuminating given concepts. We include the outdoor as well as indoor examples to further analyze the versatility of *LumiCtrl*. Though, *LumiCtrl* currently provide 7 illuminants, however, we show in Figure 7b that the user can easily interpolate between the two learned illuminants. In this way, user can synthesize their concepts in numerous intermediate illuminants between the given seven illuminants. In addition, we also ablate the controlnet in *LumiCtrl*'s pipeline, and illustrated a comparitive qualitative analysis in Fig. 8. The results show that *LumiCtrl* struggles with preserving the structural information of the input image, and introduces artifacts in the generated images when controlnet based guidance is not integrated in the training pipeline. Whereas, this problem is mitigated, when controlnet based guidance is enabled in the *LumiCtrl* pipeline. It is important to note that this conditioning mechanism is only used while learning the new concept. *LumiCtrl* do not require ControlNet during inference to maintain the generation quality.

2.6.1. Ablation Study.

We conduct ablation study over various factors. We note that *LumiCtrl* introduces artifacts in generated images, when ControlNet based guidance is removed in the training, as shown in Fig. 8. Moreover, *LumiCtrl* generates unrealistic lighting and also affects background when λ is scaled higher in masked reconstructed loss. We quantitatively analyze the effect of λ in illumination quality in generated images. In this case, we show the results for the 4 illuminants used in the user study. The results are summarized in Table 1 which show that *LumiCtrl* achieves better illumination quality with $\lambda = 0.2$.

Table 1. Ablation study over the foreground weighting hyperparameter λ in the Masked Reconstruction Loss. Lower Angular Error (AE) and MSE indicate better illuminant accuracy; higher SSIM indicate better image fidelity. Best performance is achieved at $\lambda = 0.2$.

λ	AE ↓	MSE ↓	SSIM ↑
1.0	13.20	25.60	0.41
0.8	10.10	22.30	0.58
0.6	8.40	19.70	0.60
0.4	6.90	18.10	0.65
0.2	**4.51**	**16.80**	**0.77**
0.0	9.80	20.50	0.50

References

[1] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH*, pages 1–12, 2023. 1

[2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, pages 18392–18402, 2023. 1

[3] Emmene Ton Chien. Emmene ton chien, 2024. 1

[4] Freepik. FreePik — All-in-One AI Creative Suite. 1

[5] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023. 1

[6] KHARB. Pet food - kharb, 2024. 1

[7] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, pages 1931–1941, 2023. 1

[8] Artful Paws Photography. Artful paws photography, 2021. 1

[9] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, pages 22500–22510, 2023. 1

[10] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In *The Thirteenth International Conference on Learning Representations*, 2025. 1